

Springer Series in Statistics

Geert Molenberghs and Geert Verbeke

Models for Discrete Longitudinal Data

This book provides a comprehensive treatment on modeling approaches for non-Gaussian repeated measures, possibly subject to incompleteness. The authors begin with models for the full marginal distribution of the outcome vector. This allows model fitting to be based on maximum likelihood principles, immediately implying inferential tools for all parameters in the models. At the same time, they formulate computationally less complex alternatives, including generalized estimating equations and pseudo-likelihood methods. They then briefly introduce conditional models and move on to the random-effects family, encompassing the beta-binomial model, the probit model and, in particular the generalized linear mixed model. Several frequently used procedures for model fitting are discussed and differences between marginal models and random-effects models are given attention.

The authors consider a variety of extensions, such as models for multivariate longitudinal measurements, random-effects models with serial correlation, and mixed models with non-Gaussian random effects. They sketch the general principles for how to deal with the commonly encountered issue of incomplete longitudinal data. The authors critique frequently used methods and propose flexible and broadly valid methods instead, and they conclude with key concepts of sensitivity analysis.

Without putting too much emphasis on software, the book shows how the different approaches can be implemented within the SAS software package. The text is organized so the reader can skip the software-oriented chapters and sections without breaking the logical flow.

Geert Molenberghs is Professor of Biostatistics at the Universiteit Hasselt in Belgium and has published methodological work on surrogate markers in clinical trials, categorical data, longitudinal data analysis, and the analysis of non-response in clinical and epidemiological studies. He served as Joint Editor for *Applied Statistics* (2001–2004) and as Associate Editor for several journals, including *Biometrics* and *Biostatistics*. He was President of the International Biometric Society (2004–2005). He was elected Fellow of the American Statistical Association and received the Guy Medal in Bronze from the Royal Statistical Society.

Geert Verbeke is Professor of Biostatistics at the Biostatistical Centre of the Katholieke Universiteit Leuven in Belgium. He has published a number of methodological articles on various aspects of models for longitudinal data analyses, with particular emphasis on mixed models. Geert Verbeke is Past President of the Belgian Region of the International Biometric Society, International Program Chair for the International Biometric Conference in Montreal (2006), and Joint Editor of the *Journal of the Royal Statistical Society, Series A* (2005–2008). He has served as Associate Editor for several journals including *Biometrics* and *Applied Statistics*.

The authors also wrote a monograph on linear mixed models for longitudinal data (Springer, 2000) and received the American Statistical Association's Excellence in Continuing Education Award, based on short courses on longitudinal and incomplete data at the Joint Statistical Meetings of 2002 and 2004.

ISBN 978-0-387-25144-8



springeronline.com



Preface	vii
Acknowledgments	ix
I Introductory Material	1
1 Introduction	3
2 Motivating Studies	7
2.1 Introduction	7
2.2 The Analgesic Trial	8
2.3 The Toenail Data	8
2.4 The Fluvoxamine Trial	12
2.5 The Epilepsy Data	14
2.6 The Project on Preterm and Small for Gestational Age Infants (POPS) Study	14
2.7 National Toxicology Program Data	17
2.7.1 Ethylene Glycol	18
2.7.2 Di(2-ethylhexyl)Phthalate	18
2.7.3 Diethylene Glycol Dimethyl Ether	22
2.8 The Sports Injuries Trial	23
2.9 Age Related Macular Degeneration Trial	24

3	Generalized Linear Models	27
3.1	Introduction	27
3.2	The Exponential Family	27
3.3	The Generalized Linear Model (GLM)	28
3.4	Examples	29
3.4.1	The Linear Regression Model for Continuous Data	29
3.4.2	Logistic and Probit Regression for Binary Data	29
3.4.3	Poisson Regression for Counts	29
3.5	Maximum Likelihood Estimation and Inference	30
3.6	Logistic Regression for the Toenail Data	31
3.7	Poisson Regression for the Epilepsy Data	32
4	Linear Mixed Models for Gaussian Longitudinal Data	35
4.1	Introduction	35
4.2	Marginal Multivariate Model	36
4.3	The Linear Mixed Model	36
4.4	Estimation and Inference for the Marginal Model	39
4.5	Inference for the Random Effects	41
5	Model Families	45
5.1	Introduction	45
5.2	The Gaussian Case	46
5.3	Model Families in General	47
5.3.1	Marginal Models	48
5.3.2	Conditional Models	49
5.3.3	Subject-specific Models	50
5.4	Inferential Paradigms	52
II	Marginal Models	53
6	The Strength of Marginal Models	55
6.1	Introduction	55
6.2	Marginal Models in Contingency Tables	56
6.2.1	Multivariate Logistic Models	57
6.2.2	Goodman's Local Association Models	58
6.2.3	Dale's Marginal Models	59
6.2.4	A General Class of Models	61
6.3	British Occupational Status Study	62
6.4	The Caithness Data	62
6.5	Analysis of the Fluvoxamine Trial	64
6.6	Extensions	68
6.6.1	Covariates	69

6.6.2	Three-way Contingency Tables	72
6.7	Relation to Latent Continuous Densities	79
6.8	Conclusions and Perspective	80
7	Likelihood-based Marginal Models	83
7.1	Notation	84
7.2	The Bahadur Model	86
7.2.1	A General Bahadur Model Formulation	86
7.2.2	The Bahadur Model for Clustered Data	88
7.2.3	Analysis of the NTP Data	90
7.2.4	Analysis of the Fluvoxamine Trial	92
7.3	A General Framework for Fully Specified Marginal Models	93
7.3.1	Univariate Link Functions	94
7.3.2	Higher-order Link Functions	94
7.4	Maximum Likelihood Estimation	99
7.5	An Influenza Study	99
7.5.1	The Cross-over Study	100
7.5.2	The Longitudinal Study	101
7.6	The Multivariate Probit Model	102
7.6.1	Probit Models	103
7.6.2	Tetrachoric and Polychoric Correlation	104
7.6.3	The Univariate Probit Model	105
7.6.4	The Bivariate Probit Model	106
7.6.5	Ordered Categorical Outcomes	110
7.6.6	The Multivariate Probit Model	112
7.7	The Dale Model	113
7.7.1	Two Binary Responses	113
7.7.2	The Bivariate Dale Model	115
7.7.3	Some Properties of the Bivariate Dale Model	117
7.7.4	The Multivariate Plackett Distribution	117
7.7.5	The Multivariate Dale Model	117
7.7.6	Maximum Likelihood Estimation	119
7.7.7	The BIRNH Study	119
7.8	Hybrid Marginal-conditional Specification	122
7.8.1	A Mixed Marginal-conditional Model	123
7.8.2	Categorical Outcomes	126
7.9	A Cross-over Trial: An Example in Primary Dysmenorrhoea	127
7.9.1	Analyzing Cross-over Data	128
7.9.2	Analysis of the Primary Dysmenorrhoea Data	130
7.10	Multivariate Analysis of the POPS Data	131
7.11	Longitudinal Analysis of the Fluvoxamine Study	134
7.12	Appendix: Maximum Likelihood Estimation	136

7.12.1	Score Equations and Maximization	136
7.12.2	Newton-Raphson Algorithm with Cumulative Counts	139
7.12.3	Determining the Joint Probabilities	140
7.13	Appendix: The Multivariate Plackett Distribution	142
7.14	Appendix: Maximum Likelihood Estimation for the Dale Model	147
8	Generalized Estimating Equations	151
8.1	Introduction	151
8.2	Standard GEE Theory	153
8.3	Alternative GEE Methods	161
8.4	Prentice's GEE Method	162
8.5	Second-order Generalized Estimating Equations (GEE2)	164
8.6	GEE with Odds Ratios and Alternating Logistic Re- gression	165
8.7	GEE2 Based on a Hybrid Marginal-conditional Model	168
8.8	A Method Based on Linearization	169
8.9	Analysis of the NTP Data	170
8.10	The Heatshock Study	174
8.11	The Sports Injuries Trial	181
8.11.1	Longitudinal Analysis	181
8.11.2	A Bivariate Longitudinal Analysis	186
9	Pseudo-Likelihood	189
9.1	Introduction	189
9.2	Pseudo-Likelihood: Definition and Asymptotic Prop- erties	190
9.2.1	Definition	190
9.2.2	Consistency and Asymptotic Normality	191
9.3	Pseudo-Likelihood Inference	192
9.3.1	Wald Statistic	193
9.3.2	Pseudo-Score Statistics	193
9.3.3	Pseudo-Likelihood Ratio Statistics	194
9.4	Marginal Pseudo-Likelihood	195
9.4.1	Definition of Marginal Pseudo-Likelihood	195
9.4.2	A Generalized Linear Model Representation	198
9.5	Comparison with Generalized Estimating Equations	199
9.6	Analysis of NTP Data	200
10	Fitting Marginal Models with SAS	203
10.1	Introduction	203
10.2	The Toenail Data	203

10.3	GEE1 with Correlations	204
10.3.1	The SAS Program	205
10.3.2	The SAS Output	206
10.4	Alternating Logistic Regressions	212
10.5	A Method Based on Linearization	215
10.5.1	The SAS Program for the GLIMMIX Macro . . .	215
10.5.2	The SAS Output from the GLIMMIX Macro . . .	216
10.5.3	The Program for the SAS Procedure GLIM- MIX	218
10.5.4	Output from the GLIMMIX Procedure	218
10.6	Programs for the NTP Data	219
10.7	Alternative Software Tools	221

III Conditional Models 223

11	Conditional Models	225
11.1	Introduction	225
11.2	Conditional Models	226
11.2.1	A Pure Multivariate Setting	227
11.2.2	A Single Repeated Outcome	229
11.2.3	Repeated Multivariate Outcomes	230
11.3	Marginal <i>versus</i> Conditional Models	233
11.4	Analysis of the NTP Data	234
11.5	Transition Models	236
11.5.1	Analysis of the Toenail Data	238
11.5.2	Fitting Transition Models in SAS	240
12	Pseudo-Likelihood	243
12.1	Introduction	243
12.2	Pseudo-Likelihood for a Single Repeated Binary Out- come	244
12.3	Pseudo-Likelihood for a Multivariate Repeated Bi- nary Outcome	245
12.4	Analysis of the NTP Data	246
12.4.1	Parameter Estimation	247
12.4.2	Inference and Model Selection	249

IV Subject-specific Models 255

13	From Subject-specific to Random-effects Models	257
13.1	Introduction	257
13.2	General Model Formulation	257
13.3	Three Ways to Handle Subject-specific Parameters	258

13.3.1	Treated as Fixed Unknown Parameters	258
13.3.2	Conditional Inference	258
13.3.3	Random-effects Approach	259
13.4	Random-effects Models: Special Cases	260
13.4.1	The Linear Mixed Model	260
13.4.2	The Beta-binomial Model	260
13.4.3	The Probit-normal Model	262
13.4.4	The Generalized Linear Mixed Model	262
13.4.5	The Hierarchical Generalized Linear Model	263
14	The Generalized Linear Mixed Model (GLMM)	265
14.1	Introduction	265
14.2	Model Formulation and Approaches to Estimation	265
14.2.1	Model Formulation	265
14.2.2	Bayesian Approach to Model Fitting	266
14.2.3	Maximum Likelihood Estimation	266
14.2.4	Empirical Bayes Estimation	268
14.3	Estimation: Approximation of the Integrand	268
14.4	Estimation: Approximation of the Data	269
14.4.1	Penalized Quasi-Likelihood (PQL)	270
14.4.2	Marginal Quasi-Likelihood (MQL)	270
14.4.3	Discussion and Extensions	271
14.5	Estimation: Approximation of the Integral	273
14.5.1	Gaussian Quadrature	274
14.5.2	Adaptive Gaussian Quadrature	275
14.6	Inference in Generalized Linear Mixed Models	276
14.7	Analyzing the NTP Data	277
14.8	Analyzing the Toenail Data	278
15	Fitting Generalized Linear Mixed Models with SAS	281
15.1	Introduction	281
15.2	The GLIMMIX Procedure for Quasi-Likelihood	282
15.2.1	The SAS Program	283
15.2.2	The SAS Output	284
15.3	The GLIMMIX Macro for Quasi-Likelihood	287
15.3.1	The SAS Program	288
15.3.2	Selected SAS Output	289
15.4	The NLMIXED Procedure for Numerical Quadrature	290
15.4.1	The SAS Program	290
15.4.2	The SAS Output	293
15.5	Alternative Software Tools	296
16	Marginal <i>versus</i> Random-effects Models	297
16.1	Introduction	297

16.2	Example: The Toenail Data	297
16.3	Parameter Interpretation	298
16.4	Toenail Data: Marginal <i>versus</i> Mixed Models	301
16.5	Analysis of the NTP Data	304
V	Case Studies and Extensions	307
17	The Analgesic Trial	309
17.1	Introduction	309
17.2	Marginal Analyses of the Analgesic Trial	310
17.3	Random-effects Analyses of the Analgesic Trial	314
17.4	Comparing Marginal and Random-effects Analyses	317
17.5	Programs for the Analgesic Trial	318
17.5.1	Marginal Models with SAS	318
17.5.2	Random-effects Models with SAS	320
17.5.3	MIXOR	321
17.5.4	MLwiN	323
18	Ordinal Data	325
18.1	Regression Models for Ordinal Data	326
18.1.1	The Fluvoxamine Trial	328
18.2	Marginal Models for Repeated Ordinal Data	329
18.3	Random-effects Models for Repeated Ordinal Data	331
18.4	Ordinal Analysis of the Analgesic Trial	332
18.5	Programs for the Analgesic Trial	334
19	The Epilepsy Data	337
19.1	Introduction	337
19.2	A Marginal GEE Analysis	337
19.3	A Generalized Linear Mixed Model	340
19.4	Marginalizing the Mixed Model	342
20	Non-linear Models	347
20.1	Introduction	347
20.2	Univariate Non-linear Models	349
20.3	The Indomethacin Study: Non-hierarchical Analysis	351
20.4	Non-linear Models for Longitudinal Data	355
20.5	Non-linear Mixed Models	357
20.6	The Orange Tree Data	358
20.7	Pharmacokinetic and Pharmacodynamic Models	360
20.7.1	Hierarchical Analysis of the Indomethacin Study	361
20.7.2	Pharmacokinetic Modeling and the Theophylline Data	363

20.7.3	Pharmacodynamic Data	367
20.8	The Songbird Data	368
20.8.1	Introduction	368
20.8.2	A Non-linear Mixed-effects Model	370
20.8.3	Analysis of SI at RA	371
20.8.4	Model Strategies for HVC	372
20.8.5	Analysis of SI at HVC	374
20.9	Discrete Outcomes	376
20.9.1	Analysis of the NTP Data	377
20.10	Hypothesis Testing and Non-linear Models	379
20.11	Flexible Functions	379
20.11.1	Random Smoothing Splines	381
20.11.2	Analysis of the Analgesic Trial	383
20.12	Using SAS for Non-linear Mixed-effects Models	384
20.12.1	SAS Program for the Orange Tree Data Analysis	384
20.12.2	SAS Programs for the Indomethacin Analyses	385
20.12.3	SAS Programs for the Theophylline Analyses	386
20.12.4	SAS Program for the Songbird Data	387
20.12.5	SAS Program for the NTP Data	388
20.12.6	SAS Program for the Random Smoothing Spline Model	388
21	Pseudo-Likelihood for a Hierarchical Model	393
21.1	Introduction	393
21.2	Pseudo-Likelihood Estimation	394
21.3	Two Binary Endpoints	397
21.4	A Meta-analysis of Trials in Schizophrenic Subjects	401
21.5	Concluding Remarks	403
22	Random-effects Models with Serial Correlation	405
22.1	Introduction	405
22.2	A Multilevel Probit Model with Autocorrelation	406
22.3	Parameter Estimation for the Multilevel Probit Model	408
22.4	A Generalized Linear Mixed Model with Autocorre- lation	410
22.5	A Meta-analysis of Trials in Schizophrenic Subjects	412
22.6	SAS Code for Random-effects Models with Autocor- relation	415
22.7	Concluding Remarks	417
23	Non-Gaussian Random Effects	419
23.1	Introduction	419

23.2	The Heterogeneity Model	421
23.3	Estimation and Inference	423
23.4	Empirical Bayes Estimation and Classification	427
23.5	The Verbal Aggression Data	428
23.6	Concluding Remarks	435
24	Joint Continuous and Discrete Responses	437
24.1	Introduction	437
24.2	A Continuous and a Binary Endpoint	439
24.2.1	A Probit-normal Formulation	439
24.2.2	A Plackett-Dale Formulation	441
24.2.3	A Generalized Linear Mixed Model Formulation	442
24.3	Hierarchical Joint Models	445
24.3.1	Two-stage Analysis	445
24.3.2	Fully Hierarchical Modeling	446
24.4	Age Related Macular Degeneration Trial	448
24.4.1	Bivariate Marginal Analyses	448
24.4.2	Bivariate Random-effects Analyses	452
24.4.3	Hierarchical Analyses	453
24.5	Joint Models in SAS	455
24.6	Concluding Remarks	464
25	High-dimensional Joint Models	467
25.1	Introduction	467
25.2	Joint Mixed Model	469
25.3	Model Fitting and Inference	471
25.3.1	Pairwise Fitting	471
25.3.2	Inference for Ψ	472
25.3.3	Combining Information: Inference for Ψ^*	473
25.4	A Study in Psycho-Cognitive Functioning	473
VI	Missing Data	479
26	Missing Data Concepts	481
26.1	Introduction	481
26.2	A Formal Taxonomy	482
26.2.1	Missing Data Frameworks	484
26.2.2	Missing Data Mechanisms	485
26.2.3	Ignorability	487
27	Simple Methods, Direct Likelihood, and WGEE	489
27.1	Introduction	489
27.2	Longitudinal Analysis or Not?	490

27.3	Simple Methods	491
27.4	Bias in LOCF, CC, and Ignorable Likelihood	495
27.5	Weighted Generalized Estimating Equations	498
27.6	The Depression Trial	499
27.6.1	The Data	499
27.6.2	Marginal Models	501
27.6.3	Random-effects Models	502
27.7	Age Related Macular Degeneration Trial	503
27.8	The Analgesic Trial	507
28	Multiple Imputation and the EM Algorithm	511
28.1	Introduction	511
28.2	Multiple Imputation	511
28.2.1	Theoretical Justification	512
28.2.2	Pooling Information	513
28.2.3	Hypothesis Testing	514
28.2.4	Efficiency	514
28.2.5	Imputation Mechanisms	515
28.3	The Expectation-Maximization Algorithm	516
28.3.1	The Algorithm	517
28.3.2	Missing Information	518
28.3.3	Rate of Convergence	519
28.3.4	EM Acceleration	520
28.3.5	Calculation of Precision Estimates	520
28.3.6	A Simple Illustration	521
28.4	Which Method to Use?	526
28.5	Age Related Macular Degeneration Study	527
28.6	Concluding Remarks	529
29	Selection Models	531
29.1	Introduction	531
29.2	An MNAR Dale Model	532
29.2.1	Likelihood Function	533
29.2.2	Maximization Using the EM Algorithm	535
29.2.3	Analysis of the Fluvoxamine Data	537
29.3	A Model for Non-monotone Missingness	543
29.3.1	Analysis of the Fluvoxamine Data	546
29.4	Concluding Remarks	552
30	Pattern-mixture Models	555
30.1	Introduction	555
30.2	Pattern-mixture Modeling Approach	556
30.3	Identifying Restriction Strategies	557
30.3.1	How to Use Restrictions?	560

30.4	A Unifying Framework for Selection and Pattern-mixture Models	561
30.5	Selection Models <i>versus</i> Pattern-mixture Models	563
30.5.1	Selection Models	564
30.5.2	Pattern-mixture Models	565
30.5.3	Identifying Restrictions	566
30.5.4	Precision Estimation with Pattern-mixture Models	566
30.6	Analysis of the Fluvoxamine Data	567
30.6.1	Selection Modeling	568
30.6.2	Pattern-mixture Modeling	569
30.6.3	Comparison	572
30.7	Concluding Remarks	572
31	Sensitivity Analysis	575
31.1	Introduction	575
31.2	Sensitivity Analysis for Selection Models	576
31.3	A Local Influence Approach for Ordinal Data with Dropout	578
31.3.1	General Principles	578
31.3.2	Analysis of the Fluvoxamine Data	581
31.4	A Local Influence Approach for Incomplete Binary Data	585
31.4.1	General Principles	585
31.4.2	Analysis of the Fluvoxamine Data	586
31.5	Interval of Ignorance	590
31.5.1	General Principle	591
31.5.2	Sensitivity Parameter Approach	593
31.5.3	Models for Monotone Patterns and a Bernoulli Experiment	594
31.5.4	Analysis of the Fluvoxamine Data	599
31.6	Sensitivity Analysis and Pattern-mixture Models	604
31.7	Concluding Remarks	605
32	Incomplete Data and SAS	607
32.1	Introduction	607
32.2	Complete Case Analysis	607
32.3	Last Observation Carried Forward	609
32.4	Direct Likelihood	611
32.5	Weighted Estimating Equations (WGEE)	613
32.6	Multiple Imputation	618
32.6.1	The MI Procedure for the Imputation Task	618
32.6.2	The Analysis Task	624
32.6.3	The Inference Task	629

32.6.4	The MI Procedure to Create Monotone Missingness	633
32.7	The EM Algorithm	633
32.8	MNAR Models and Sensitivity Analysis Tools	635
	References	637
	Index	671